DOCUMENT RESUME

ED 467 808 TM 034 350

AUTHOR	Wang, Xiang Bo; Pan, WeiQin; Harris, Vincent
TITLE	Computerized Adaptive Testing Simulations Using Real Test Taker Responses. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-96-06
PUB DATE	1999-03-00
NOTE	12p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Ability; *Adaptive Testing; College Entrance Examinations; *College Students; *Computer Assisted Testing; Estimation (Mathematics); Higher Education; Law Schools; Maximum Likelihood Statistics; *Responses; *Simulation
IDENTIFIERS	Accuracy; *Law School Admission Test

ABSTRACT

A considerable amount of data on computerized adaptive testing (CAT) has been conducted using simulated data. However, most researchers would agree that simulations may not fully reflect the reality of examinee performance on a test. This study used maximum likelihood procedures to investigate the accuracy and efficiency of examinee ability recovery using real test taker responses (n=969) to 127 items from the Law School Admission Test (LSAT). It was hoped that through control of item parameters and examinee abilities based on a realistic test, insights would be gained on practical issues involved in an operative computerized adaptive LSAT. The 127 items were found to be sufficient to conduct the CAT sessions for virtually all test takers at the 3 accuracy levels. The highest accuracy level seemed to be necessary to recover the original paper-and-pencil ability estimates of the test takers. The average number of items used in the CAT sessions ranged from 13 for the lowest accuracy level to 37 for the highest accuracy level. This study represented the first CAT simulation that used real test taker responses. (Contains 4 figures, 6 tables, and 11 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made

ED 467 808

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. VASELECK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Computerized Adaptive Testing Simulations Using Real Test Taker Responses

Xiang Bo Wang, WeiQin Pan, and Vincent Harris
Law School Admission Council

Law School Admission Council
Computerized Testing Report 96-06
March 1999

A Publication of the Law School Admission Council

TM034350

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

LSAT®; *The Official LSAT PrepTest®*; *LSAT: The Official TriplePrep®*; and the Law Services logo are registered marks of the Law School Admission Council, Inc. Law School forum is a service mark of the Law School Admission Council, Inc. *LSAT: The Official TriplePrep Plus*; *The Whole Law School Package*; *The Official Guide to U.S. Law Schools*, and ***LSACD*** are trademarks of the Law School Admission Council, Inc.

Copyright© 1999 by Law School Admission Council, Inc.

All rights reserved. This book may not be reproduced or transmitted, in whole or in part, by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary 1

Abstract 1

Introduction 1

Research Method, Questions, and Data 2

Analyses and Results 3
Summarizing Paper and Pencil Item and Test Taker Ability Parameter Estimates 3
CAT Simulations 4

Conclusion 7

References 8

Executive Summary

Simulations have played an important role in the research of computerized adaptive testing (CAT), ranging from the study of estimation accuracy, item exposure, test security, to item differential functioning. However, most researchers would agree that few simulations, regardless of how well conducted, can fully capture the psychological and behavioral reality of examinee performance on a test. This is especially true when it comes to simulating individual test taker responses.

In order to maintain the full information embodied in test taker responses, this study used the real responses of 969 Law School Admission Test (LSAT) takers to 127 items to simulate a simple CAT. (A practical constraint of using real data is the limited sizes of item and test taker pools.) The objective of this study was to use the original ability estimates of these test takers obtained from their paper-and-pencil (P & P) LSAT as targets to see how well simple CAT sessions could recover them. Three basic research questions were investigated: (1) Can the limited number of items (127) of one P & P test be adequate for a CAT session? (2) How accurately can such a CAT recover individual test taker abilities? (3) How many items are required by CAT sessions with three accuracy levels—standard error of measurement (SEM) equal to 0.3162, 0.265, and 0.173, which are equivalent to the classical reliabilities of 0.90, 0.93 and 0.97, respectively? Note that the reliability of the LSAT is normally about 0.93.

Three major findings have been obtained. First, the 127 items have been shown to be sufficient to conduct simple CAT sessions for virtually all test takers at the three accuracy levels. Only two test takers' CAT sessions could not be completed at the highest accuracy level. The second finding is that given the current item pool composition, the highest accuracy level seemed to be necessary to recover the original P & P ability estimates of the test takers. The third major finding is that the average number of items used in CAT sessions increased from 13 items for the lowest accuracy level, to 37 items for the highest accuracy level. The overall significance of this study is that it is the first CAT simulation that used real test taker responses.

Abstract

Considerable amount of research in computerized adaptive testing (CAT) has been conducted using data of simulated examinees, item responses, and item parameters. However, most researchers would agree that few simulations, regardless of how well it is conducted, can fully reflect the reality of examinee performance on a test. Based on maximum likelihood procedures, this research investigates the accuracy and efficiency of examinee ability recovery using real test taker responses to the Law School Admission Test (LSAT) items. It is hoped that through the control of item parameters and examinee abilities based on a realistic test, insights will be gained on various practical issues concerning an operational computerized adaptive LSAT in the future.

Introduction

Simulations have played an omnipresent role in the research of computerized adaptive testing (CAT). Recent examples include the studies on item selection criteria by Veerkamp and Berger (1994); impact of bias of maximum likelihood estimation by Kinsbury (1996); controlling item exposure rates by Stocking (1993); CAT item differential functioning by Zwick, Thayer, and Wingersky (1995); balancing item information, content, and exposure by Luecht, Nungester, and Hadadi (1996); and the possibilities of constructing CATs to parallel conventional paper-and-pencil (P & P) tests by Davey (1996). However, most researchers would agree that few simulations, regardless of how well conducted, can fully reflect the psychological and behavioral reality of test taker performance on a test. This is especially true when it comes to simulating individual test taker responses.

One typical stochastic method to simulate test takers' right or wrong responses is to compare random uniform probabilities with the probabilities yielded by some statistical models, such as the three-parameter logistic (3PL) item response theory (IRT) model. While such a simulation method serves some purposes, its adequacy can be contended when the goal is to simulate test taker responses for a particular test, such as the

Law School Admission Test (LSAT). This is because test takers' responses on a test are products of many factors working together, such as test takers' educational experiences, cognitive processes, test behavior, specific time limits, and psychological and physical conditions at the time of the test, to name just a few. However, the statistical item response simulation method described above and IRT on which most CAT simulations are based only assume that a unidimensional ability trait is solely responsible for test taker responses. It is clear that item responses generated by a statistical simulation method are not a holistic representation of real test taker responses.

It is a fact that simulating any of the above-mentioned real-life factors that impact on test taker responses is extremely difficult, if not impossible. For example, speededness and guessing toward the end of a test or section—common test behaviors—are still very poorly simulated. In an attempt to avoid the shortcomings of the stochastic simulation method, this study used real test taker responses, obtained from an administration of the LSAT, to produce more realistic CAT simulations.

Research Method, Questions, and Data

The LSAT is a paper-and-pencil (P & P) test designed to assess the essential skills of reading comprehension, logical reasoning, and analytical reasoning (LSAC, 1996) for law school studies. During an LSAT administration, test takers are typically administered approximately 127 items divided into four operational sections and one variable section. The four operational sections consist of one Analytical Reasoning (about 25 items), two Logical Reasoning (about 50 items) and one Reading Comprehension (about 26 items). The variable section can be any one of the three section types and is designed to pretest new test items or to preequate new test forms.

The real responses used in this study come from the 969 LSAT test takers who answered the 101 operational items, and a same set of 26 variable reading comprehension items of a disclosed LSAT administration. In other words, the study is based on 969 test takers' responses to 127 items. This research design thus uses real responses as CAT responses during CAT simulation, instead of using stochastically generated responses. These real responses are the holistic embodiment of the test takers in terms of their educational experiences, cognitive processes, test behavior, and psychological and physical conditions at the time of the test.

Three basic research questions will be investigated in this study: (1) Can the limited number of items (127) of one P & P test be adequate for a CAT session? (2) How accurately can CAT recover individual test taker θ's? (3) How many items are required by CAT sessions with standard error of measurement (SEM) equal to 0.3162, 0.265, and 0.173, which are equivalent to the classical reliabilities of 0.90, 0.93 and 0.97, respectively? No content constraints were imposed in this study for basically two reasons. First, as this study is the first of its kind, some baseline statistical information has to be obtained before content constraints confound it. Second, although content constraints are well established for the P&P LSAT, no content constraints have been instituted for the CAT version of the LSAT. Effects of the content constraints will be studied in a future research project.

This study employs the 3PL model for item response modeling, maximum likelihood method for test taker ability (θ) estimation, and maximum information for item selection.

Regarding the CAT simulation algorithm, all CAT sessions start from the same place—with an item which has the maximum item information near -0.25. No mathematical estimation of ability is used for the first five items. If a test taker provides a mixture of correct and incorrect responses for the first five items, the test taker's ability is increased by 0.5 for each correct response, and decreased by 0.25 for each incorrect response. Newton-Raphson estimation and item selection routine are employed from the sixth item on. If a test taker answers the five items, either all correctly or all incorrectly, the test taker's ability estimate is increased or decreased five times, each time at an increment of 0.5. Such a process is extended beyond the first five items until the test taker provides an opposite answer.

A CAT is considered successfully completed only when both a Newton-Raphson converged ability estimate θ and one of the three preestablished SEM levels are achieved. To reiterate, the three predetermined SEM

levels 0.3162, 0.265, and 0.173 are equivalent to the classical reliabilities of 0.90, 0.93, and 0.97, respectively. Please note in this study that the SEM is computed by:

$$SEM = \sqrt{\frac{1}{\sum_{j=1}^{n} I_j(\hat{\theta}_i)}}$$

where n stands for the total number of items used in a CAT session; j, the item index; i, the examine index; θ, test taker ability estimate; and I, item information.

Analyses and Results

Summarizing Paper and Pencil Item and Test Taker Ability Parameter Estimates

Before presenting CAT simulation results, it is of interest to show the distributions of item and ability parameter estimates as derived from their original P & P test. Table 1 summarizes the distributions of the *a*, *b*, and *c* parameter estimates of the 127 items used in this study. The *a*, *b*, and *c* parameter estimates represent item discrimination, difficulty, and a pseudo-guessing factor, respectively. It can be seen from Table 1 that most items have moderate discrimination, centering around 0.70 with the minimum being 0.32 and the maximum 1.51. The difficulties of the majority of the items are situated around the middle difficulty range, 0.17, where the least difficult item is -2.54 and the most difficult item is 3.65. The pseudo-guessing parameter of the items is very low, around 0.17.

TABLE 1

Item parameter descriptive statistics

Parameter	N	Mean	STD	Minimum	Medium	Maximum	Range
a	127	0.70	0.21	0.32	0.66	1.51	1.18
b	127	0.17	1.12	-2.54	0.08	3.65	6.19
c	127	0.17	0.11	0.00	0.17	0.50	0.50

According to Figure 1, the ability distribution of the 969 test takers is approximately normal with a slight skew toward the low ability range. According to line 1 of Table 2, the mean, minimum, and maximum of the ability distribution are 0.07, -3.70 and 3.31, respectively. Note that the distribution of test taker abilities is of special interest here because it is the target that the CAT simulation will attempt to recover.

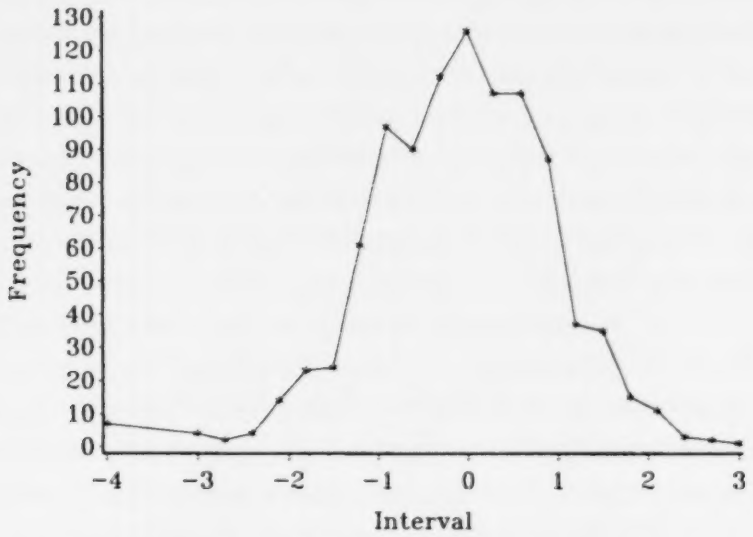

FIGURE 1. *Distribution of test taker ability estimate*

TABLE 2

Test taker ability descriptive statistics

	N	Mean	STD	Minimum	Medium	Maximum	Range
Original	969	0.07	0.98	-3.70	0.10	3.31	7.01
CAT-1	968	0.07	1.15	-3.82	0.06	3.96	7.79
CAT-2	968	0.10	1.10	-3.85	0.08	3.73	7.57
CAT-3	966	0.09	1.00	-3.25	0.06	3.28	6.53

Note. CAT-1: using real response and SEM = 0.316 equivalent to $\rho = 0.90$.
CAT-2: using real response and SEM = 0.265 equivalent to $\rho = 0.93$.
CAT-3: using real response and SEM = 0.173 equivalent to $\rho = 0.97$.

CAT Simulations

Corresponding to the three predetermined SEM levels, three CAT simulations—CAT-1, CAT-2, and CAT-3—were carried out in this study. How well did the three CAT sessions recover the original P & P ability distribution of the test takers? Results summarized in Table 2 demonstrate that all three CAT sessions virtually replicated the original ability distribution—all the descriptive statistics—mean, standard deviation, minimum, medium, and maximum, are quite similar.

However, varying degrees of differences do exist when the ability estimates of individual test takers are compared between their original P & P and CAT ability estimates. An overview of Figures 2-4, which display the differences of ability estimates of individual test takers, reveals three trends. First, the differences between the original and CAT ability estimates of individual test takers gradually decrease as the accuracy level increases from CAT-1 to CAT-3. (That is when SEM decreases.) For example, according to Figure 2, at the SEM level of 0.316 or classical reliability of 0.90, individual test takers whose original P&P ability estimates were located around 0.00 in IRT scale, could receive CAT ability estimates ranging from about -1.00 to about 1.50, which is certainly a substantial level of variation. Yet, such differences at the same ability level of 0.00, as shown in Figure 4, were drastically reduced between -0.30 to 0.50, when the SEM level was lowered to 0.173, equivalent to the classical reliability of 0.97 in CAT-3.

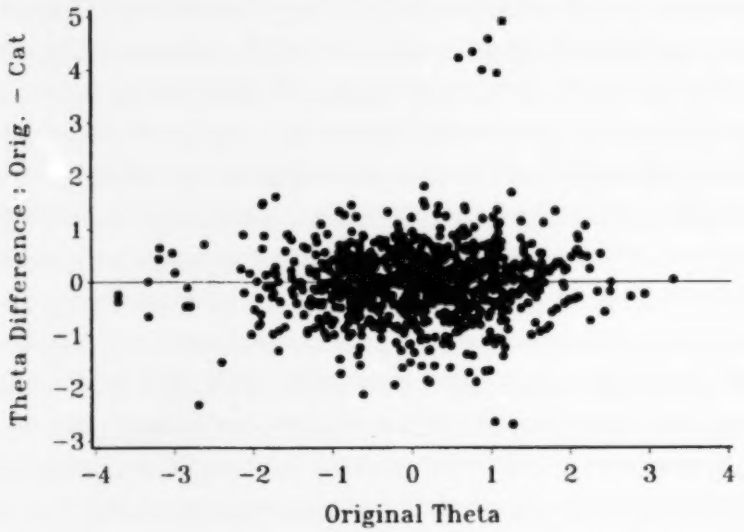

FIGURE 2. *Differences between original and CAT-1 ability estimates*

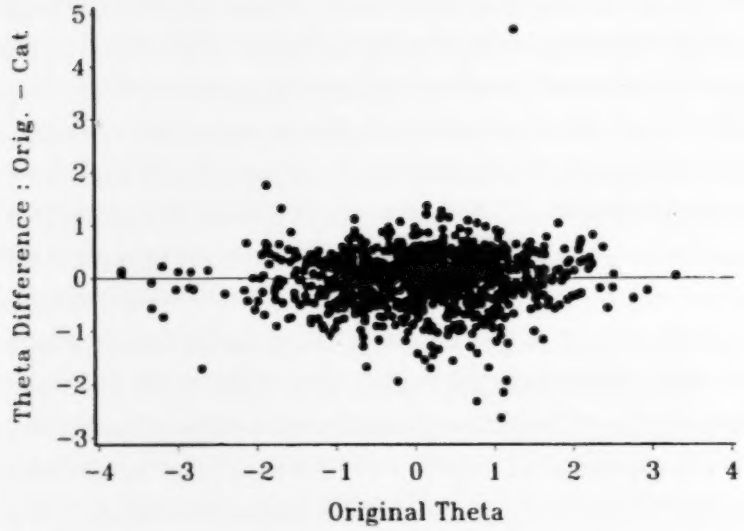

FIGURE 3. *Differences between original and CAT-2 ability estimates*

BEST COPY AVAILABLE

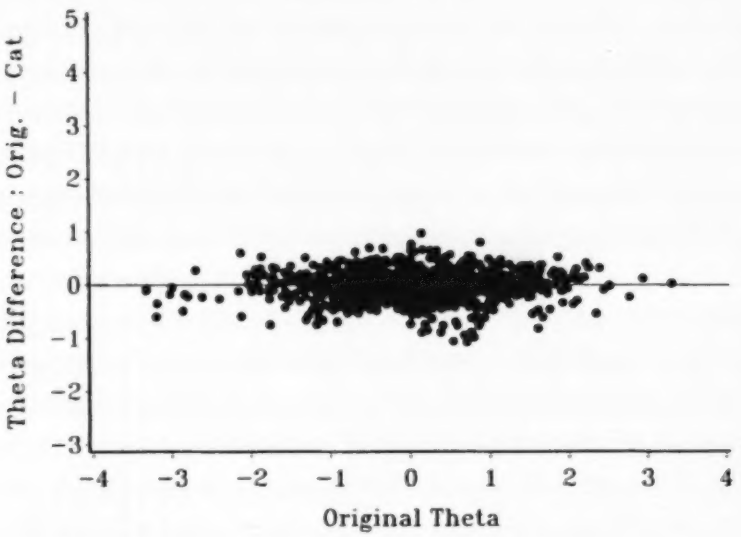

FIGURE 4. *Differences between original and CAT-3 ability estimates*

Table 3 summarizes the descriptive statistics about the improvement. For example, from CAT-1 to CAT-3, the mean absolute differences decreased from 0.48 to 0.22; the standard deviation of the differences, from 0.50 to 0.18; and the maximum differences, from 4.92 to 1.07.

TABLE 3

Descriptive statistics on the absolute differences between CAT and original ability estimates

	N	M	STD	Minimum	Medium	Maximum	Range
CAT-1 - Original	968	0.48	0.50	0.00	0.36	4.92	4.92
CAT-2 - Original	968	0.39	0.36	0.00	0.32	4.70	4.70
CAT-3 - Original	966	0.22	0.18	0.00	0.18	1.07	1.07

The second trend noted for three CAT sessions was that most differences occurred among middle-range ability. This finding is reasonable because both extremely low- and high-ability test takers tended to get items either wrong or correct, respectively. The CAT item selection process tended to have little or no effect on their ability estimates. However, since middle-ability test takers did have a number of items answered either correctly or incorrectly, the items the CAT item selection process chose, would certainly affect their ability estimates. If the CAT item selection process picked items mostly from those originally answered correctly, these middle-range ability test takers would obtain ability estimates higher than their original ones. The reverse would happen if the CAT item selection process picked items mostly from those that middle-range ability test takers originally answered incorrectly. Such an effect of item selection would be especially obvious when the accuracy level is set low (when the SEM level is large). A common strategy to battle such an effect is to increase CAT accuracy by lowering the SEM level.

How many items did the three CAT sessions use on the average? Table 4 reveals that as accuracy improved from CAT-1 to CAT-3, the average number of items also increased substantially, from about 13 items for CAT-1 sessions, to 17 for CAT-2 sessions, to 37 items for CAT-3 sessions. For some test takers in CAT-1, as few as seven items were sufficient, while for others in CAT-3, as many as 127 items, namely the entire CAT item pool in this study, were insufficient.

TABLE 4

Comparison of number of items selected by CAT sessions

	N	Mean	STD	Minimum	Medium	Maximum	Range
CAT-1	968	12.92	5.63	7	12	56	49
CAT-2	968	17.08	5.77	9	16	72	63
CAT-3	966	36.74	9.05	21	35	127	106

What kind of test takers took as many as 127 items for their CAT-3 session? Table 5 shows that four test takers used up the entire item pool—three had extremely low-ability estimates, while one had an extremely high-ability estimate. The reason these four test takers used up all the items is that the limited CAT item pool of 127 items used in this study did not have sufficient items for their extremely high- or low-ability ranges. Instead of optimal items, suboptimal items were used for their ability ranges and the predetermined accuracy level could never be reached.

TABLE 5

Summary of the test takers who took 127 items in CAT-3

Parameter	N	Mean	STD	Minimum
1	24	-3.32	-3.22	127
2	24	-2.99	-2.94	127
3	26	-3.32	-3.25	127
4	123	3.31	3.28	127

As far as item exposure is concerned, Table 6 shows that some items were used for every test taker, while other items were used a few times. The most frequently used items were those used to initiate CAT sessions, because all CAT sessions started from the same place— -0.25, as stated earlier. No specific explanation can be offered for the most infrequently used items. Unlike previous studies, no consistent correlation has been found between the item use frequencies and the magnitude of item information, probably due to the limited size of the item pool.

TABLE 6

Comparison of item use frequency

Parameter	N	Mean	STD	Minimum	Medium	Maximum	Range
CAT-1	127	98.50	137.18	2	65	968	966
CAT-2	127	130.00	136.79	7	95	968	961
CAT-3	127	281.45	128.18	19	277	968	949

Conclusion

As the first simulation project to employ real test taker responses, this study has obtained two major findings concerning realistic CAT simulations in the context of the LSAT. First, it has shown that it is possible to conduct CAT sessions using 127 items from one P & P LSAT administration, especially at the precision level of SEM equal to 0.316 or reliability of 0.90. Even at the high precision level of SEM of 0.173 or reliability level of 0.97, 127 items are sufficient except for four test takers. However, estimation difficulties did occur for test takers of extremely high or low abilities, because the current item pool does not contain enough items that are sufficient to cover the extreme ranges. The second finding is that the precision level of SEM equal to 0.173 or reliability of 0.97 seems to be necessary in order to produce CAT ability estimates that are sufficiently close to their P & P ability counterparts. This is especially true if the item pool is limited in size or short of highly informative items (items with high *a* parameters) across the entire ability range.

BEST COPY AVAILABLE

The shortcoming of this study is due to the practical limitation of real responses; only 127 items were used, not enough to provide optimal information for the entire ability range. As a result, for some test takers, fewer items could have been used if more optimal items had been available.

In order to overcome the limitation of the small number of items, the authors plan, for future research, to increase the number of items by pooling items from different LSAT administrations and linking test taker responses of similar abilities. The increased item pool will make it possible to assess the SEM of CAT ability estimates by repeated CAT sessions for each test taker through random selections of items. The SEM of CAT ability estimates is a vital mathematical property for statistical inferences. With the increased item pool and random item selection, the authors will also investigate the effects of content constraints, conditional exposure rates of items and item types, and similarities and differences of the CAT ability estimates between using real and simulated test taker responses.

The advent of CAT has offered more efficiency and accessibility to test takers. However, CAT also poses various new challenges, one of which is the adequacy of simulations that lead to actual operational CAT programs. Real test taker responses automatically contain all the cognitive, psychological, and behavioral factors that underlie test taker responses but are missing in simulated data. It is fair to say that virtually all existing CAT programs that are currently in operation have been based on simulated responses. They may have defects, some of which may be very serious and can be partially attributed to unrealistic simulations during their developmental stages. The solution to avoid such potential problems may very well lie in research designs that employ realistic simulations. It is hoped that this study will stimulate more research that employs more realistic simulations, ultimately producing more stable CAT programs.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring a test taker's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Davey, Tim (1996, April). *Constructing adaptive tests to parallel conventional programs.* Paper presented at the annual conference of the National Council on Measurement in Education: New York.

Kingsbury, G. (1996, April). *The impact of bias in maximum-likelihood on the accuracy of scores from computerized adaptive tests.* Paper presented at the annual conference of the National Council on Measurement in Education: New York.

Luecht, R., Nungester, R., & Hadadi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure.* Paper presented at the annual conference of the National Council on Measurement in Education, New York.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31,* 251-263.

LSAT/LSAS Registration and Information Book 1996-97 (1996). Newtown, PA: Law School Admission Council.

Stocking, M. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report No. 93-2). Princeton, NJ: Education Testing Service.

Veerkamp, W., & Berger, M. (1994). *Some new item selection criteria for adaptive testing* (Research Report No. 94-6). Enschede, Netherlands: University of Twente, Faculty of Educational Science and Technology.

Wainer, H., Dorans, N., Flaugher, R., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (1990) *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ward, W., Potenza, M., & Stocking, M. (1997). *Simulating the use of disclosed items in computerized adaptive testing* (Research Report No. 97-10). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32* (4), 341-363.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").